Data QUEST
Data Quality Testing – DQe Tools
2/28/17

Kari A. Stephens, PhD
Assistant Professor, Psychiatry & Behavioral Sciences
Adjunct Assistant Professor, Biomedical Informatics & Medical Education
WWAMI region Practice & Research Network

- ~58 Primary care WWAMI clinics
- ~20 data connected clinics
- CHCs and RHCs
- Underserved populations
- Many serving rural populations
- Collaboration with national network of practice based research networks
- Data QUEST represents over 250,000 patients
  
  https://dataquest.iths.org/
  https://github.com/WWAMI-DataQuest
Data QUEST

• 20 data-connected clinics in the WPRN
• Represents over 250,000 patients

An electronic health data-sharing architecture across community-based primary care practices in the WPRN
Data QUEST: Improving Health in Rural Populations

funded by NIH, AHRQ, CDC, PCORI, AHRQ, CMMS, and industry

Current Clinical Research Trials

- Team-based Safe Opioid Prescribing – dissemination trial across 6 regional primary care practices (AHRQ)
- Integrating Behavioral Health and Primary Care – large national pragmatic trial across 40 national primary care practices (PCORI)

Network Participation

- PCORNNet’s Patient-Centered Scalable National Network for Effectiveness Research (pSCANNER) (PCORI)
- Clinical Trials Network: Pacific Northwest Node (NIH/NIDA)
- Accelerating Change and Transformation in Organization and Networks III (ACTION III) partnership, The Quality Commons (AHRQ)
- WWAMI Practice Transformation Network (CMS)
- Diabetes Prevention Registry (CDC)
- Northwest Pharmacogenomic Research Network (NIH/NIGMS)
- DARTNet Practice Benchmarking Registry (industry)
- MOSAIC: Meaningful Outcomes and Science to Advance Innovations Center of Excellence (AHRQ)
Common Tables – OMOP V.4

• Care Site
  • Sites at each organization
• Condition Occurrence
  • Encounter associated diagnoses
  • Problem list diagnoses
• Drug Exposure
  • Medications
• Location
  • Patient and site addresses

• Observation
  • Vitals and Labs
  • Past medical history
  • Family history
• Person
  • Patient demographics
• Procedure Occurrence
  • Encounter associated procedures
  • CPT codes
• Visit Occurrence
  • Appointments
  • Encounters
310,604 patients in the person table

- 102,330 (33%) at Organization B
- 45,685 (15%) at Organization C
- 27,577 (9%) at Organization N
- 36,001 (12%) at Organization P
- 99,011 (32%) at Organization Y

10M encounters
Measuring Data Quality
A new framework...

Operationalizing the framework into: 5 conceptual tests and 17 discrete tests across:

- **Completeness**
  - Are the data present?

- **Conformance**
  - Are the data standardized and formatted?

- **Plausibility**
  - Are the data believable?


<table>
<thead>
<tr>
<th>TEST ID</th>
<th>DOMAIN</th>
<th>TEST</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>COMPLETENESS</td>
<td>Number of Tables Received, Number of Observations, Flag Indicator for the table having actual data</td>
</tr>
<tr>
<td>C2</td>
<td>COMPLETENESS</td>
<td>GENDER completeness (denominator and proportion with valid data)</td>
</tr>
<tr>
<td>C3</td>
<td>COMPLETENESS</td>
<td>AGE/DOB completeness (denominator and proportion with valid data)</td>
</tr>
<tr>
<td>C4</td>
<td>COMPLETENESS</td>
<td>VITALS completeness (denominator and proportion with valid data): Height, Weight, SBP, DBP</td>
</tr>
<tr>
<td>C5</td>
<td>COMPLETENESS</td>
<td>LABS completeness (denominator and proportion with valid data): A1c, HDL, LDL, Triglycerides, Total cholesterol</td>
</tr>
<tr>
<td>F1</td>
<td>FIDELITY</td>
<td>Check that primary and foreign keys relate properly; High Priority: Person_ID, Visit_Occurrence_ID</td>
</tr>
<tr>
<td>F2</td>
<td>FIDELITY</td>
<td>Duplicate patient check in the patient table (Find the same patient with a different patient ID using full name, dob, and gender)</td>
</tr>
<tr>
<td>F3</td>
<td>FIDELITY</td>
<td>Visualize codes/values entered for DEMOGRAPHICS (Gender, Race, Ethnicity)</td>
</tr>
<tr>
<td>F4</td>
<td>FIDELITY</td>
<td>Visualize YEAR OF BIRTH to help identify errors or missing cohorts</td>
</tr>
<tr>
<td>P1</td>
<td>PLAUSIBILITY</td>
<td>Comparison of new load to old load (Number of observations, Number of unique patients, Number of tables with rows)</td>
</tr>
<tr>
<td>P2</td>
<td>PLAUSIBILITY</td>
<td>Review of minimum and maximum dates for tables with key dates; High Priority: Visit_Occurrence table</td>
</tr>
<tr>
<td>P3</td>
<td>PLAUSIBILITY</td>
<td>How many patients have a year of birth after their visit dates?</td>
</tr>
<tr>
<td>P4</td>
<td>PLAUSIBILITY</td>
<td>Check that certain observation types fall into specific ranges</td>
</tr>
<tr>
<td>P5</td>
<td>PLAUSIBILITY</td>
<td>Visualize number of visits in a year or across years</td>
</tr>
<tr>
<td>P6</td>
<td>PLAUSIBILITY</td>
<td>Visualize type of visit in a year or across years</td>
</tr>
<tr>
<td>P7</td>
<td>PLAUSIBILITY</td>
<td>Volume Check: Proportion of patients with visit data and select observation types</td>
</tr>
<tr>
<td>P8</td>
<td>PLAUSIBILITY</td>
<td>Logical Constraints Check</td>
</tr>
</tbody>
</table>
DQe Tool Architecture

DQe-c
modular tool developed in R statistical language for assessing completeness in EHR data repositories

DQe-v
interactive interface powered by the shiny package version 0.13.0 in R
Operationalizing use of DQe tools for data quality testing

* Data QUEST
* DARTNet Institute
Create a dataset of data quality related measures (for instance, visits per year) sorted by measure, organization, and year.

Read the data and run the DQe-v R script.

Review HTML output for data quality issues related to plausibility across multiple organizations.

Review HTML output of the DQe-c Add-On for data quality issues related to completeness, fidelity, and plausibility ACROSS multiple organizations.

Run R script for the DQe-c Add-On against the individual organization report files generated during the main DQe-c report process.

Review HTML output of individual DQe-c reports for data quality issues related to completeness, fidelity, and plausibility.

Run the DQe-c R script against the CDM for each organization individually.
The network’s table schemas and key relationships
• Color coated to display “missingness”
Completeness example:
Number of primary keys for available tables over time
Completeness example:
Detailing columns with proportion of missingness (null vs. blank)
Fidelity example:
Detailing totals of key overlap across core tables

Count of Unique person_id in Tables with person_id

Count of Unique care_site_id in Tables with care_site_id
Completeness/Fidelity example:
Percent of patients missing specific key clinical indicators

- gender: 0.01%
- encounter: 33.89%
- weight: 42.83%
- diagnosis: 44.00%
- race: 44.45%
- height: 45.66%
- ethnicity: 48.97%
- medication: 55.96%
- smoking: 69.43%
- alcohol: 100%
Completeness/Fidelity example across sites:
Percent of patients missing specific key clinical indicators
Completeness example across sites/clinics:
Percent of patients missing in columns across sites

<table>
<thead>
<tr>
<th>person</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>year_of_birth</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>race_source_value</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>race_concept_id</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>provider_id</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>person_source_value</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>person_id</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>month_of_birth</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>location_id</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gender_source_value</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gender_concept_id</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ethnicity_source_value</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ethnicity_concept_id</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>day_of_birth</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>care_site_id</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| procedure_occurrence    |          |          |          |          |          |          |
| visit_occurrence_id      |          |          |          |          |          |          |
| relevant_condition_concept_id |      |          |          |          |          |          |
| procedure_type_concept_id |          |          |          |          |          |          |
| procedure_source_value   |          |          |          |          |          |          |
| procedure_occurrence_id  |          |          |          |          |          |          |
| procedure_date           |          |          |          |          |          |          |
| procedure_concept_id     |          |          |          |          |          |          |
| person_id                |          |          |          |          |          |          |
| associated_provider_id   |          |          |          |          |          |          |
Plausability example across sites/clinics:
# of Hemoglobin A1c’s per year per diabetes patient with 1+ visit

Zoom to 2015-16
Next Steps

• Finalize SOP manual for DQe
• Iterate and refining functionality in DQe-v
• Create standard report of data quality findings
• Add new tests as needed...

Thank you!

Contact: Kari Stephens
kstephen@uw.edu

https://dataquest.iths.org/

https://github.com/WWAMI-DataQuest